データから得られる情報・知識・知恵

北海道大学　情報基盤センター
先端データ科学研究室　水田 正弘

１．はじめに

用語としての「データ」と「情報」の区別をあいまいに扱うことがよくあります。また、「知識」と「知恵」についても同様です。データから有益な知見を得るためには、これらの用語を区別することが大切だと考えられます。大きな流れとしては、「データ」から「情報」を作り、さらに「知識」を獲得します。「知恵」については、この流れの潤滑油としての側面と、流れの到達点としての側面があります。

医療・ビジネス・教育研究のどの分野でも、データに基づく判断が重視されています。非常に少数のデータであれば、数値を眺めるだけで、知りたいことが見えてくるかもしれません。しかし、近年、我々が直面するデータの量は、増加する一方であり、かつ、そのデータ構造も、複雑化しています。このために、データを扱う方法を、単なる技術ではなく、情報・知識・知恵の観点から理解する必要があります。

本報告では、統計学やデータ解析の数多くの手法、さらには S-PLUS をはじめとする統計パッケージを上記の観点から考えたいと思います。さらに、本テーマに関係したデータ解析に関する最新の研究状況を紹介いたします。

２．データと情報、そして知識へ

新聞などに、「本世論調査のデータによると」と書かれることは多いのですが、「本世論調査の情報によると」はほとんど使われません。また、実験で得られた数値のことを「実験データ」とよび、「実験情報」とはよびません。

マクドノウは、情報を「特定の状況における価値が評価されたデータ」と定義しています。すなわち、「データ」とは単なる数値などの集まりであり、何らかの価値観・基準・目的を意識することにより「情報」となります。データ解析の基本的な役割は、データから情報を取り出すことであります。

例えば、「平成 21 年度全国学力・学習状況調査」、いわゆる学力テストが平成 21 年 4 月 21 日に実施されました。これについてある県が、「データ開示」しました。新聞などによると「国語Ａ、Ｂと算数・数学Ａ、Ｂの平均正答数と平均正答率」を市町村別および学校別の一覧として開示したとのことです。開示したこと自体の評価はここでは議論をひかえます。また、多くの官公庁で「データ開示」または「データ公表」を実施しています。これらの「データ」を「情報」にするには、なんらかの価値を設定し、評価することになります。

この種のデータのもっとも基本的な集計は平均です。平均を用いて、グループ（市町村、

学校など）に順位をつけて並べ替えることで、一つの「情報」となります。しかし、この「情報」を生成する基準となった価値・評価はかなり単純なものと言えます。分散（または標準偏差）を使うと、ある仮定のもとで平均の差に意味があるかを調べることができます。回帰分析、判別分析、主成分分析、クラスター分析など比較的複雑な手法を用いると別の切り口から「情報」を生成することになります。つまり、一つの「データ」から複数の「情報」が作成されるとこになります。重要な点はどのような価値・評価を設定しているかです。

「情報」から普遍化できた事項を「知識」とよびます。基本的で汎用性の高い「知識」は教科書や百科事典に書かれるかもしれません。人間のパターン認識能力は、大変優れているので、いくつかの「情報」から「仮説」を作り出すことができます。この「仮説」を検証することができれば「知識」となります。この検証において統計学が有力な方法であることは言うまでもありません。

３．データ解析における知恵、データ解析により得られる知恵

本やウェッブページで使われる表現に「おばあちゃんの知恵袋」があります。たまに「おじいちゃんの知恵袋」もあります。この「おばあちゃん」や「おじいちゃん」は、経験豊かな人の象徴と考えられます。「ファスナーに布がからんだときの対処法」や「飲み終わったお茶パックの有効利用法」など、学校では習わない、しかし、便利な「知恵」が多数、掲載されています。

統計学やデータ解析は、大学をはじめとする学校で講述されます。体系的な講義を目指すため、確率論や数学に基づく理論的な内容が中心の場合が少なくないと思われます。大学の教師としては、これは当然のことであり、理論的な基礎を学ぶことの重要性を強く主張したいと思います。しかし、より実践的なデータの解析においては、個々の理論（分布論、検定、アルゴリズム）や手法（差の検定法、分散分析、相関ルール）に関する基本的な理解のもとに、どのような手順で解析を実施すべきかが本質的になってきます。

統計相談を受ける時、相談者は何をしたいかを聞きだすことが第一歩です。それに従い、どのようなデータを収集し、解析目的に応じた手法やモデルを考え、実際にコンピュータで実行する方法までを相談することになります。これは、「知恵を貸す」ことです。

分野が限られていると、かなり多くの実践的な知恵が存在します。例えば、経済関係におけるデータを扱った門倉貴史(2006)では、多くの間違った解析例を示し、著者の解釈を加えています。

データ解析自体の知恵について紹介しましたが、データ解析により知恵を得ることも可能です。初めに紹介した通り、「データ」から「情報」を生成し、「知識」を見つけ出すプロセスがあります。どの分野でも、「情報」や「知識」を総合的に判断して、行動を決定することは典型的なパターンだと思われます。このことも、重要な「知恵」と言えます。例えば、経営に対する知恵、営業の対する知恵、診断に対する知恵です。もちろん、これま

での「経験」が重要です。認知心理学などいくつかの分野で、熟達化という用語を使います。「学習者は熟達の過程で，先人の知恵の集積するコミュニティ（実践家の集団）で行われる文化によって組織化された実践活動に参加し，多くのことを学ぶ」と人工知能学辞典に説明されています。

このような知恵の体系化は困難だと思いますが、データ解析または統計パッケージにより支援は可能だと思います。

４．新しいデータ解析法

これまでの多くのデータ解析手法では、解析対象が個体（例えば、各患者）であり、その個体をいくつかの数値で表現されたものをデータとして想定しています。それに対し、集団や概念なども含めた広い対象を解析する方法としてシンボリックデータ解析法が提唱されました(Diday *et al.*, 2007)。これは、解析対象を、その帰属を定義する方法と、属する集合で示されているコンセプト(Concept)としています。各コンセプトは多様な方法で記述されます。当然、通常の数値や多次元ベクトルで記述することもありますが、それ以外に、集合、区間値、分布値などがあります。コンセプトからなるデータをシンボリックデータとよびます。

このようなシンボリックデータに対する解析手法として、従来の回帰分析、主成分分析、判別分析、クラスター分析などの拡張が提案されています。シンボリックデータ解析は、Fuzzyや確信度なども含めた広いアプローチをとっています。

５．おわりに

データ・情報・知識、さらには知恵という流れは、統計学やデータ解析に限定すべきものではなく、ビジネス、教育、実生活において重要なものです。本報告では、データ解析の立場から扱っていきましたが、それぞれの分野において固有の流れがあると思います。それをサポートするツールとしてデータ解析システムは有効だと思います。データ解析の研究は、多くの方が想像する以上のスピードで進んでいます。数理システムユーザーコンファレンスなど多くの機会を通じて情報交換ができればと思っております。

参考文献

Diday, Edwin and Noirhomme-Fraiture, Monique (2007), Symbolic Data Analysis and the SODAS Software, Wiley.

水田正弘・山本義郎・南 弘征・田澤 司(2005), S-PLUSによるデータマイニング入門, 森北出版.

門倉貴史(2006), 統計数字を疑う なぜ実感とずれるのか？ 光文社新書.

人工知能学会編(2008), デジタル人工知能学辞典, 共立出版.